Machine Intelligence Research
www.mi-research.net
DOI: 10.1007/s11633-022-1330-7

Towards Interpretable Defense Against Adversarial Attacks via Causal Inference

Min Ren[1,2] Yun-Long Wang[2] Zhao-Feng He[3]

[1]University of Chinese Academy of Sciences, Beijing 100190, China

[2]Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[3]Laboratory of Visual Computing and Intelligent System, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: Deep learning-based models are vulnerable to adversarial attacks. Defense against adversarial attacks is essential for sensitive and safety-critical scenarios. However, deep learning methods still lack effective and efficient defense mechanisms against adversarial attacks. Most of the existing methods are just stopgaps for specific adversarial samples. The main obstacle is that how adversarial samples fool the deep learning models is still unclear. The underlying working mechanism of adversarial samples has not been well explored, and it is the bottleneck of adversarial attack defense. In this paper, we build a causal model to interpret the generation and performance of adversarial samples. The self-attention/transformer is adopted as a powerful tool in this causal model. Compared to existing methods, causality enables us to analyze adversarial samples more naturally and intrinsically. Based on this causal model, the working mechanism of adversarial samples is revealed, and instructive analysis is provided. Then, we propose simple and effective adversarial sample detection and recognition methods according to the revealed working mechanism. The causal insights enable us to detect and recognize adversarial samples without any extra model or training. Extensive experiments are conducted to demonstrate the effectiveness of the proposed methods. Our methods outperform the state-of-the-art defense methods under various adversarial attacks.

Keywords: Adversarial sample, adversarial defense, causal inference, interpretable machine learning, transformers.

Citation: M. Ren, Y. L. Wang, Z. F. He. Towards interpretable defense against adversarial attacks via causal inference. *Machine Intelligence Research*. http://doi.org/10.1007/s11633-022-1330-7

1 Introduction

Deep learning methods open a new era of artificial intelligence. In the field of computer vision, deep learning methods have achieved great success in image classification[1–7], object detection[8, 9], and image segmentation[10, 11]. Deep neural networks showcase the powerful capability to perform a nonlinear mapping from raw data to high-level features. However, adversarial samples cast a shadow over the notable success of deep learning. The "powerful" deep learning modules are vulnerable to various adversarial attacking algorithms[12–14]. Using well-crafted perturbations, attackers can undermine predictions from state-of-the-art models, even though the perturbations cannot be spotted by humans. This problem prevents the application of deep methods in sensitive and safety-critical scenarios[15–18]. Hence, defense against adversarial attacks is of considerable concern and has become an essential research topic.

Numerous studies on defense against adversarial attacks have been reported. However, it is unclear how adversarial samples fool deep learning models. The underlying working mechanism of adversarial samples deserves more exploration and study. Hence, most existing methods are just stopgaps for specific adversarial samples. For example, adversarial training, which introduces adversarial samples into the training process, is widely popular as a defense method. However, the generalization capability of adversarial training-based methods is quite limited, especially for unseen attacks.

In order to defend against adversarial attacks, it is necessary to reveal the working mechanism of adversarial samples. In this paper, we adopt causal inference to explore the working mechanism of adversarial samples as shown in Fig. 1. Compared to the methodology based on statistics, causal inference models the relationship between variates more naturally and intrinsically. A causal model is established to describe the generation and performance of adversarial samples. The causal model enables us to estimate the causal effects between the outputs of the deep neural network and the subregions of adversarial samples, which cannot be realized by data-driv-

Research Article
Manuscript received February 22, 2022; accepted April 13, 2022
Recommended by Associate Editor Min-Ling Zhang
Colored figures are available in the online version at https://link.springer.com/journal/11633

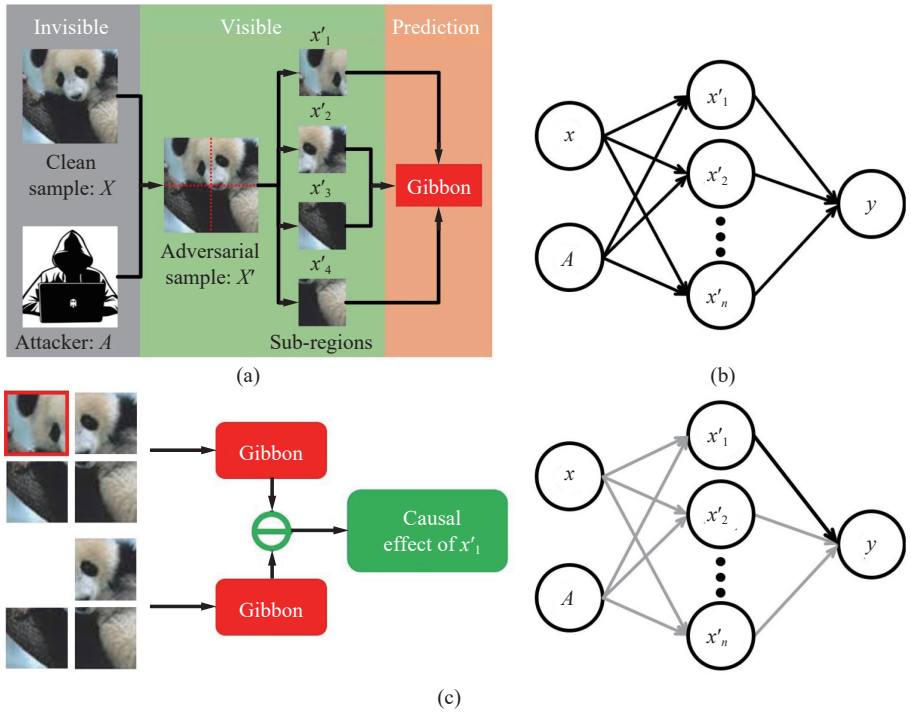

Fig. 1 Estimating causal effects between the outputs and the subregions of adversarial samples: (a) Generation and performance of adversarial samples; (b) Causal graph of the adversarial sample; (c) Causal effect of a subregion on the prediction is estimated by the intervention treatment.

en/statistical methodology. Hence, the tampered predictions can be attributed to the subregions, which means that it is possible to interpret the adversarial sample and reveal its working mechanism.

Because of its special construction, vision transformer (ViT)[19] is adopted to establish the causal model. ViT is a self-attention-based deep learning model. It originates from transformers, which is designed to process variable-length sequential data. To handle images, ViT divides images into non-overlapping patches for use as input. Hence, ViT can handle the variable-length sequential input, i.e., variable number of image patches. This characteristic of ViT enables us to perform the intervention treatment on patches and attribute the outputs to subregions of images.

According to the discoveries of the working mechanism of adversarial samples, we propose simple and effective strategies for defense against adversarial attacks. The causal analysis reveals that the causal effect of different subregions of an adversarial sample can be inconsistent, even opposing. Usually, only a small part of the adversarial sample plays a decisive role in fooling the recognition model. Hence, adversarial attacks can be defended according to the inconsistency of the adversarial samples. Moreover, the causal effect of subregions at different scales can differ, which indicates that it is helpful to fuse the multi-scale inconsistencies for defense against adversarial attacks. Note that the recognition model does not need to be retrained using adversarial samples and that there are few performance costs for clean samples. It is almost free to detect adversarial attacks and improve the robustness of the recognition model because the working mechanism of the adversarial samples is revealed by our causal model.

The main contributions of this paper can be summarized as follows:

1) We build a causal model to interpret the generation and performance of adversarial samples. The causal model enables us to estimate the causality between the outputs of the deep neural network and the subregions of the input samples.

2) Based on causal inference, the working mechanism of the adversarial samples is revealed. The causal effect of different subregions of an adversarial sample can be inconsistent, even opposing. Usually, only a small portion of the adversarial sample plays a decisive role in fooling the recognition model.

3) According to these discoveries, we propose simple and effective strategies for defense against adversarial attacks. These strategies enable us to detect and recognize adversarial samples without additional models or training.

The remainder of this paper is organized as follows: Section 2 presents a brief literature review on the related work. The causal model of adversarial samples is detailed in Section 3. The causal model is utilized to analyze adversarial samples in Section 4. The proposed strategy for adversarial attack detection and quantitative evaluations are presented in Section 5. Then, the strategy for adversarial sample recognition and the quantitative evaluations are presented in Section 6. Finally, the conclusions of this paper are summarized in Section 7.

2 Related work

In this section, we provide a brief overview of the research on adversarial samples as references, the self-attention-based deep learning models designed for images, and the applications of causal inference in the field of computer vision.

2.1 Adversarial attack

The adversarial attack method for computer vision tasks is a research hotspot. Szegedy et al.[12] first demonstrated that deep neural networks are vulnerable to adversarial perturbations. Subsequently, many adversarial attacking methods have been proposed. Goodfellow et al.[13] proposed an efficient single-step attack method named fast gradient sign method (FGSM), which is a gradient-based method. DeepFool[14] seeks to find the nearest decision boundary to confuse the model. C&W[20] was proposed to solve the joint optimization of the objective function and the perturbation scale. Projected gradient descent (PGD)[21] iteratively applies FGSM. Generalization of the adversarial perturbations has also been reported. Attack methods based on universal adversarial perturbation are proposed in [22–25].

Recently, some researchers have proposed subregion-based adversarial attack methods. Differently from the methods meantioned above, which falsify the whole image, subregion-based adversarial attack methods manipulate subregions of the image to fool the recognition model. There is usually no constraint on the scale of adversarial perturbations. An extreme case confuses the recognition model by changing only a single pixel in the image[26]. The subregion-based adversarial attack is easy to realize physically through methods such as stickers/patches on traffic signs, patterned eyeglass frames, and 3D printed objects[27–31].

These methods reduce the costs of adversarial attacks and increase the challenge for recognition systems in the real world.

2.2 Defense against adversarial attacks

1) Adversarial attack detection

A family of defense strategies is adversarial attack detection, which attempts to distinguish between benign and adversarial samples. For example, Feinman et al.[32] developed a logistic regression-based (LR) adversarial example detector that uses kernel density and Bayesian uncertainty features. Ma et al.[33] estimated a local intrinsic dimensionality (LID) score at each neural network layer and characterize key properties of the adversarial subspace. Recently, Yu et al.[34] proposed an algorithm with two key steps: i) the application of Gaussian noise in the input example, and ii) the use of the number of steps required to change the classification of the example (from benign to adversarial and vice versa) as a distance metric to detect adversarial attacks. Another strategy is to use the nearest neighbor algorithm for adversarial attack detection. The deep k-nearest neighbors (DkNN)[35] method uses a k-nearest neighbor model at every layer of the network to assess whether the input example is adversarial. Nearest neighbors, especially those that do not belong to the majority class, are used for this determination. Lee et al.[36] proposed a Mahalanobis distance-based method that models the distribution of samples in each class independently.

However, for learning-based methods, the generalization ability is limited. The kNN-based methods suffer from high computational complexity. Based on the new viewpoint of causality, our method is simple and effective. The recognition model does not need to be retrained by adversarial samples, which is computationally efficient.

2) Adversarial sample recognition

Another line of work on defense is robust recognition. Existing adversarial sample recognition methods can be roughly classified into two categories. The methods in the first category aim to improve the robustness of neural networks against adversarial examples. Methods in the second category attempt to erase adversarial perturbations from the samples before feeding them to the target model.

A common strategy of the first category is to train networks with adversarial examples[13, 37–39]. Various learning strategies have been proposed to improve robustness against gradient-based attacks. Ross and Doshi-Velez[40] trained the model while regularizing the input gradients. Cazenavette et al.[41] attempted to improve the adversarial robustness of CNNs by reframing each layer as a sparse coding model. Network distillation[42], the region-based classifier[43], the generative model[44, 45], and self-supervised learning[46] have also been adopted to improve the robustness of the models.

Some other methods are designed to remove the adversarial perturbations before the recognition model. Das et al.[47] sought to remove the perturbations using joint photographic experts group (JPEG) compression. PixelCNN[48] was used to transform the adversarial examples to clean images in [49, 50]. Moosavi-Dezfooli et al.[51] and Sun et al.[52] adopted sparse coding to reconstruct patches of images. Self-supervised learning was adopted to remove the adversarial noise in class activation feature space in [53].

Most of the existing adversarial sample recognition methods are costly in real-world applications where new adversarial attack methods are constantly emerging, since they are just stopgaps for specific adversarial attacks. The underlying working mechanism of adversarial samples has not been well interpreted. As a result, the generalization ability of these methods to unseen adversarial attacks is limited.

2.3 Causal inference in computer vision

Recently, causal inference[54, 55] has been introduced in various computer vision fields, including feature learning[56, 57], few-shot classification[58], long-tailed recognition[59], semantic segmentation[60], and visual question answering[61].

In the field of multi-instance learning (MIL), where an object is represented as a bag of instances, some researchers have also studied the idea of estimating the causal effect between an instance (patch) to the bag (image) label[62, 63]. The causal structure of an image is separated as an accumulation of the patches′ causal effects in these methods due to the task configuration and the limitations of CNNs. These methods assume that the patches′ causal effects can be simply combined by an OR operation. However, this assumption cannot be applied to adversarial samples. As we will demonstrate in Section 4.3, there is no simple accumulation relationship between the causal effects of different scale subregions. The causal structure of the adversarial sample is much more complex. Hence, MIL methods are not appropriate for adversarial samples.

The working mechanism of adversarial samples deserves further exploration and study, as it is the bottleneck for research on adversarial samples. We are the pioneer in examining adversarial samples from the causal viewpoint. Based on causal inference, we provide a simple and effective method for defense against adversarial attacks.

2.4 Self-attention/Transformers in computer vision

Inspired by the success of self-attention layers and transformer architectures in natural language processing (NLP), self-attention structures have been introduced into the field of computer vision. Some works employed self-attention layers to replace some or all of the spatial convolution layers in the popular ResNet[64–66]. Other researchers attempted to augment a standard CNN architecture with self-attention layers or transformers[67, 68]. ViT[19] directly applies a transformer architecture to non-overlapping image patches for image classification. The pioneering work of ViT and its follow-ups[69–71] have achieved impressive performance in image classification compared to convolutional networks.

The characteristics of ViT enable us to realize the causal inference on patches and attribute the recognition result to subregions of images.

3 Causal model of adversarial samples

In this section, we build a causal model to describe the generation and performance of adversarial samples. Based on this causal model, we adopt ViT to realize causal effect estimation and attribute the outputs of recognition to the subregions of adversarial samples. Accordingly, we can interpret the working mechanism of adversarial samples.

3.1 Notations

We denote an image sample using x. An image sample can be divided into subregions: $\{x_1, x_2, \cdots, x_n\}$, where n is the number of subregions. The adversarial attack algorithm generates an adversarial sample:

$$x' = \mathcal{A}(x) \tag{1}$$

where $\mathcal{A}(\cdot)$ is the adversarial attack algorithm and x' is the adversarial sample generated from x. The adversarial sample x' can also be divided into subregions in the same way: $\{x'_1, x'_2, \cdots, x'_n\}$. The prediction of the recognition model is denoted by y, i.e., the predicted category to which the input sample belongs.

3.2 Causal graph of adversarial samples

First, we construct a causal graph[54, 72] of the adversarial sample and the prediction, as shown in Fig. 2. This causal graph is a directed acyclic graph used to indicate how variables interact with each other through causal links. The direction of a link indicates the direction from cause to result. The adversarial sample is divided into subregions in the causal graph, and each subregion has a causal effect on the prediction. Note that it is an extreme case, all subregions have causal effects on the prediction. It simplifies the causal model, since it is difficult, if not impossible, to point out which subregion of a sample has a noteworthy causal effect on the prediction before the specific analysis. Furthermore, it does not interfere with the following quantitative analysis. A subregion without a causal effect on the prediction can be described as having a causal effect of 0.

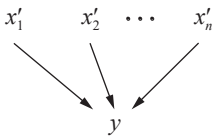

Fig. 2 Causal graph of the adversarial sample and the prediction of the recognition model. The adversarial sample is divided into subregions in the causal graph, and each subregion has a causal effect on the prediction.

Then, the adversarial attack algorithm can be introduced into this causal graph. The adversarial sample is generated by the attack algorithm. Hence, the attack algorithm is one of the causes of the adversarial sample. Accordingly, the causal graph is expanded, as shown in Fig. 3.

The clean image sample is another cause of the ad-

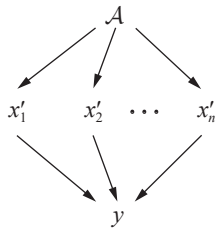

Fig. 3 The adversarial attack algorithm is introduced into the causal graph.

versarial sample. As shown in Fig. 4 (a), a subregion of the adversarial sample is generated from the corresponding subregion of the clean sample by adding adversarial perturbations. Hence, there is a causal effect from the subregion of the clean sample to the corresponding subregion of the adversarial sample. Accordingly, the causal graph can be expanded, as shown in Fig. 5.

Meanwhile, the other subregions of the clean sample also have a causal effect on this subregion of the adversarial sample, as shown in Fig. 4 (b). All parts of the clean sample are taken into consideration by the attack algorithm during the generation of the adversarial sample. Hence, there are causal effects from all subregions of the clean sample on this subregion of the adversarial sample. Accordingly, the causal graph can be expanded, as shown in Fig. 6.

In addition, the subregions of the clean sample are not independent. There are common causes for these subregions since they are split from the same image. Despite the common causes being unobservable, they can be abstracted into two hidden variates: the category-specific hidden cause, denoted as Z, and the category-invariant hidden cause, denoted as R. Hence, the causal graph is expanded as shown in Fig. 7.

This causal graph is the causal model of the adversarial samples. It illustrates the generation and influences on the prediction of the adversarial sample. Based on this causal model, we can interpret the working mechanism of the adversarial samples through causal effect estimation.

3.3 Causal effect estimation

As shown in the causal graph in Fig. 7, there are many confounders between the subregion of the adversarial sample and the prediction. A confounder is a variable that influences both causes to result and creates a spurious statistical correlation. For example, x_1 is a confounder of x'_1 and y:

$$x'_1 \longleftarrow x_1 \longrightarrow x'_2 \longrightarrow y$$

The attack algorithm, the subregions of the clean sample, and their common causes are confounders of the subregion of adversarial samples and the prediction. Meanwhile, these variates in the causal graph are unobservable in the test configuration. The impact of these confounders cannot be removed from a statistical perspective, e.g., the data-driven machine learning methodology. Hence, it is difficult to directly estimate the causal effect between the subregion of the adversarial sample and the prediction.

In order to overcome this problem, a counterfactual sample is generated to realize an intervention treatment for the performance of the adversarial sample. The causal effect of a subregion can be estimated by comparing the prediction of the adversarial sample and that of its corresponding counterfactual sample. Without loss of generality, we use the causal effect between x'_i and the prediction as an example. We introduce a variable f_i as a flag of x'_i. When $f_i = 1$, x'_i is input into the recognition model normally. When $f_i = 0$, x'_i is removed from the input of the recognition model, and the input sample becomes a counterfactual sample. As a result, the causal effect between a subregion and the prediction can be estimated by comparing the prediction of the adversarial sample and that of its corresponding counterfactual sample:

$$\phi_i = y_{f_i=1} - y_{f_i=0} \tag{2}$$

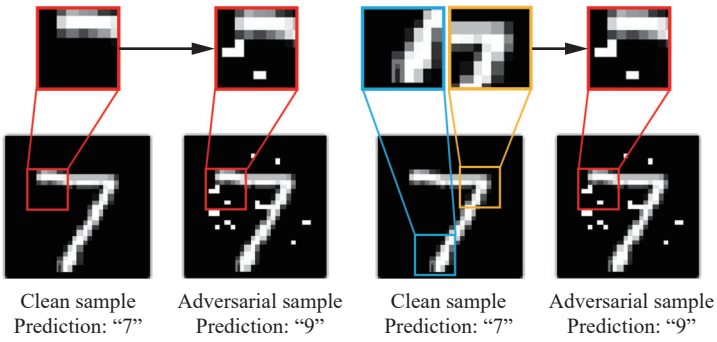

Fig. 4 An example image from MNIST. The prediction is changed from "7" to "9" by the adversarial sample: (a) There is a causal effect from the subregion of the clean sample to the corresponding subregion of the adversarial sample; (b) There are also causal effects from the subregions in different locations of the clean sample.

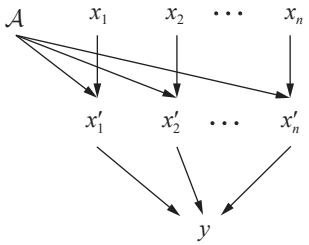

Fig. 5 Causal effect from the subregion of the clean sample on the corresponding subregion of the adversarial sample is introduced into the causal graph.

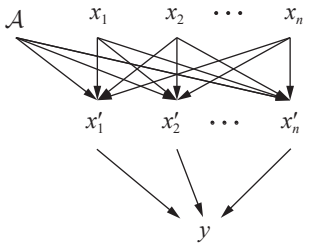

Fig. 6 Causal effects of the other subregions of the clean sample on the subregion of the adversarial sample

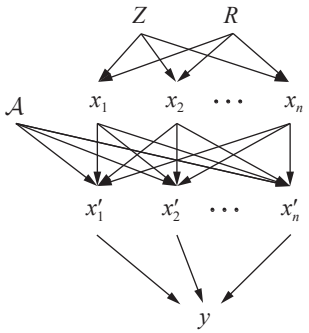

Fig. 7 Complete causal graph

where ϕ_i is the estimated causal effect of the i-th subregion.

Fortunately, the ViT is a proper framework for the proposed intervention treatment. Transformer, which is a self-attention-based architecture, has become the model of choice in NLP[73]. It is designed for variable-length sequential inputs. To address image samples, ViT splits an image into non-overlapping patches and provides the sequence of linear embeddings of these patches as an input to a transformer. The image patches are treated in the same way as tokens (words) in an NLP application. Hence, ViT can handle the variable-length sequential input, i.e., variable number of image patches. A patch of the image, as an input token, can be conveniently removed according to the corresponding flag, and the other patches will not be disturbed. On the other hand, the convolutional neural network (CNN) is unsuitable for intervention treatment. The subregion of the input image can only be replaced rather than removed, which would introduce unwanted interference. Therefore, it is appropriate to utilize ViT as a recognition model to realize the proposed intervention treatment.

Based on the proposed methodology of causal effect estimation, we can attribute the prediction of the recognition model to subregions of the adversarial sample. The pseudo-code for causal effect estimation is shown in Algorithm 1. This methodology enables us to interpret the working mechanism of adversarial attacks.

Algorithm 1. Causal effect estimation

Require: Adversarial sample: x'; a division of x': $\{x'_1, x'_2, \cdots, x'_n\}$; pre-trained ViT: $\mathcal{F}$; the predicted category: y.

Ensure: Causal effect of the subregions of x' on y: $\{\phi_1^y, \phi_2^y, \cdots, \phi_n^y\}$.

1) **for** $i = 1 \to n$ **do**
2) $f_i = 1$, set the i-th flag
3) $y_{f_i=1} \leftarrow \mathcal{F}(x'_1, x'_2, \cdots, x'_n)$
4) $X_i = 0$, set the i-th flag
5) $y_{f_i=0} \leftarrow \mathcal{F}(x'_1, \cdots, x'_{i-1}, x'_{i+1}, \cdots, x'_n)$
6) $\phi_i \leftarrow y_{f_i=1} - y_{f_i=0}$
7) **end for**
8) **return** $\{\phi_1^y, \phi_2^y, \cdots, \phi_n^y\}$

4 Causal effects of adversarial samples

In this section, we utilize the proposed methodology for causal effect estimation to explore the causal effects of adversarial samples.

In addition, the causal effects of different adversarial attack methods and different scales are also investigated.

The pre-trained ViT-Base/16, which contains 12 layers, is adopted as the recognition model. Two datasets are taken into consideration: CIFAR-10[74] and ImageNet[75], and all image samples are reshaped to 256×256.

4.1 Comparing to clean samples

To reveal the working mechanism of adversarial samples, the adversarial samples are compared with clean samples in this subsection. There are two adversarial attack configurations: targeted attack and non-targeted attack. In the targeted attack scenario, the adversary aims to induce the recognition model to give a specific label to the input sample. If there is no specific label, the attack is non-targeted, which means the adversary only wants the recognition model to predict incorrectly. Both configurations are investigated using the proposed methodology. In order to uncover the differences between the causal effects of misclassified clean samples and those of adversarial samples, the misclassified clean samples are taken separately and compared to the adversarial samples.

Projected gradient descent (PGD)[21] is adopted as the attack algorithm, which generates adversarial samples. The adversary can perturb the clean sample within a cer-

tain amount:

$$\left\|x - x'\right\|_{\infty} < \epsilon \tag{3}$$

where $\epsilon = 8/255$ for all of the adversarial samples.

Two quantitative indexes are adopted to uncover the differences between adversarial samples and clean samples. The first is the sparseness of the positive effect:

$$S = 1 - \frac{N_{positive}}{N_{total}} \tag{4}$$

where S is the sparseness of the positive effect of an input image, $N_{positive}$ is the number of patches that have positive causal effects, and N_{total} is the total number of patches. The higher the S, the patches with positive causal effects are sparser.

The second is the total variation of causal effect, which measuress the discontinuity of causal effect:

$$TV = \sum_{m,n} \sqrt{(\phi_{m+1,n} - \phi_{m,n})^2 + (\phi_{m,n+1} - \phi_{m,n})^2} \tag{5}$$

where TV is the total variation of the causal effect, (m, n) is the 2D index of patches in the input image, and $\phi_{m,n}$ is the causal effect of the patch with the 2D index (m, n).

The causal effects of examples from CIFAR-10 and ImageNet are shown in Fig. 8. The positive causal effects are denoted in red, and the negative causal effects are denoted in green. The quantitative results are shown in Table 1. mS refers to the mean of the sparseness of the positive effect, and sdS refers to the standard deviation of the sparseness. Similarly, mTV and sdTV are the mean and standard deviation of the total variation of the causal effect, respectively. "Clean Mis." refers to the misclassified clean samples. "Adv."refers to the adversarial samples.

From the visualized causal effects and the quantitative results, four points about adversarial samples can be summarized as follows:

1) Compared to clean samples, the patches with positive causal effects are sparser in adversarial samples. Only a small portion of the subregions contribute to fooling the recognition model, while the causal effects of most subregions are opposite or negligible. This phenomenon indicates that only some key subregions of the adversarial sample play a decisive role, although the whole sample is perturbed by adversarial noise.

2) The spatial continuity of causal effects of adversarial samples is lower than that of clean samples. The caus-

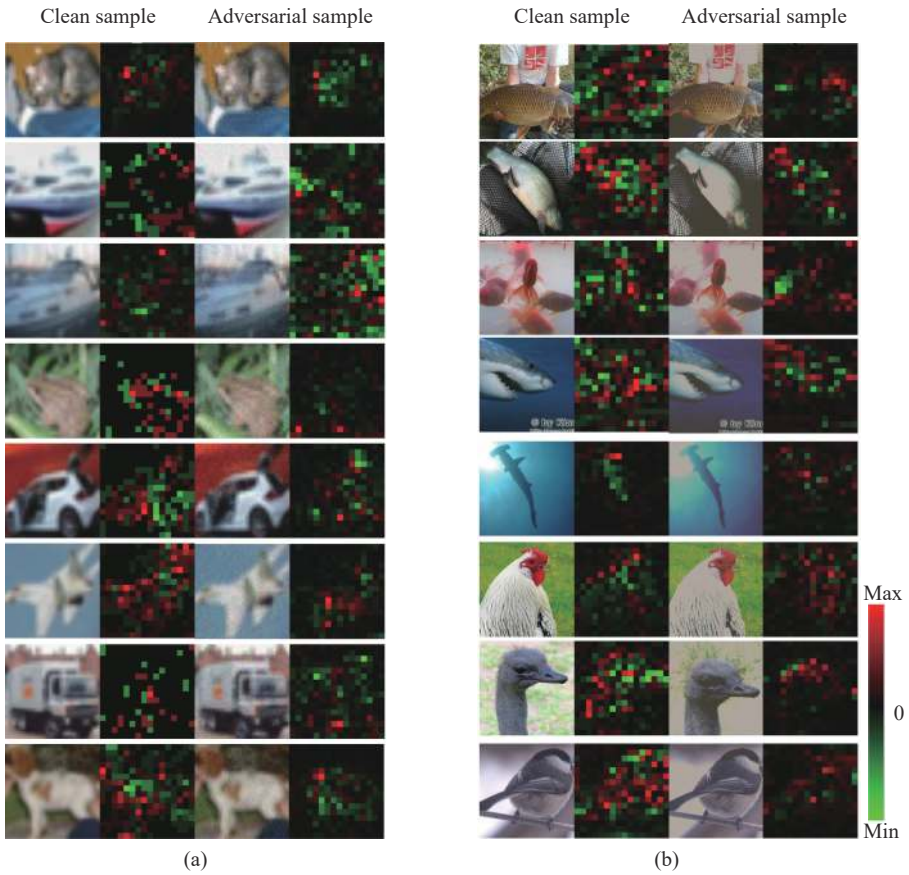

Fig. 8 Causal effects of clean and adversarial samples: (a) Causal effects of the samples in CIFAR-10; (b) Causal effects of the samples in ImageNet. The clean images are shown on the left and the adversarial samples are shown on the right.

Table 1 Quantitative results for clean samples and adversarial samples

	CIFAR-10				ImageNet			
	mS	sdS	mTV	sdTV	mS	sdS	mTV	sdTV
Clean	0.751	0.164	60.32	30.65	0.535	0.067	65.20	24.68
Clean Mis.	0.791	0.084	59.95	25.44	0.522	0.056	65.66	24.04
Adv. target	0.917	0.049	81.90	47.32	0.823	0.078	104.81	48.82
Adv. untarget	0.935	0.061	72.05	33.39	0.826	0.089	79.84	28.33

al effects of adjacent subregions within an adversarial sample are more likely to play different roles, which indicates that the causal effect of an adversarial sample is complicated.

3) Compared to the misclassified clean samples, the first two points about adversarial samples are still tenable. There are insignificant differences between correctly classified and misclassified clean samples in the sparseness of the positive effect and the spatial continuity of the causal effect. This phenomenon shows that the working mechanism of adversarial samples is quite different from clean samples, no matter whether clean samples can be correctly classified or not.

4) As we can observe from the visualized causal effects in Fig. 8, compared to the regions with less information, e.g., the plain background, the causal effects of regions with more semantic information, e.g., the foreground, are more significant. This phenomenon indicates that the adversarial attacker is more likely to tamper with the existing content of the image sample rather than to create something out of nothing.

4.2 Causal effects of different kinds of attacks

Is there a difference between the working mechanism of adversarial samples generated by different adversarial attack algorithms? In order to answer this question, four kinds of adversarial attacks are explored. We begin with one of the most basic, the fast gradient sign method (FGSM)[13]. The iterative fast gradient sign method (ITGSM)[76], is also investigated. The stronger attack methods included are the Carlini and Wagner (C & W) attack[20] and projected gradient descent (PGD)[21]. ϵ is also set to 8/255 for all adversarial samples. The two quantitative indexes, the sparseness of the positive effect and the total variation of the causal effect, are adopted to measure the differences in the causal effects.

The causal effects of examples from CIFAR-10 and ImageNet are shown in Fig. 9. The quantitative results are shown in Table 2. The analysis in the last subsection is appropriate for all adversarial attack algorithms. This observation indicates that they are common characteristics of an adversarial attack.

Meanwhile, the quantitative indexes of FGSM and ITFGSM differ slightly more than the other attack methods, since they are similar in attack mechanisms. This phenomenon shows that the proposed method captures the principal characteristics of adversarial samples.

4.3 Causal effects at different scales

What are the relationships between the causal effects of different scales? In order to answer this question, the proposed methodology is applied to three different scales: 16×16, 32×32 and 64×64. The input images are divided into patches of three sizes to investigate the causal effects at different scales.

The results on CIFAR-10 and ImageNet are shown in Fig. 10. Two points can be deduced from the results:

1) The causal effect of a large patch is not equivalent to an accumulation of small patches. There is no linear accumulation relationship between the causal effects of different scales. The causal effects could even reverse on a larger scale. This phenomenon conforms to our common sense. In many scenarios, it is impossible to recognize an object according to any portion of it. Only the entirety has effective causal effects, which means the causal effects of the entirety are not the accumulation of the portions.

2) The negative causal effects of the small scales alleviate or invert at the larger scale in most cases. This phenomenon indicates that the adversarial sample works better on a larger scale.

5 Adversarial attack detection

In Section 3, we build a causal model for adversarial samples. Based on this model, we analyze the working mechanism of adversarial attacks using causal effect estimation. This analysis is instructive for adversarial attack detection. In this section, we propose a simple and effective strategy for adversarial attack detection according to the discovery in the last section.

5.1 Detection based on semantic inconsistency

As we have mentioned in Section 4.1, the causal effects of the subregions are not consistent within an adversarial sample. Only a small portion of the subregions contribute to fooling the recognition model, while the

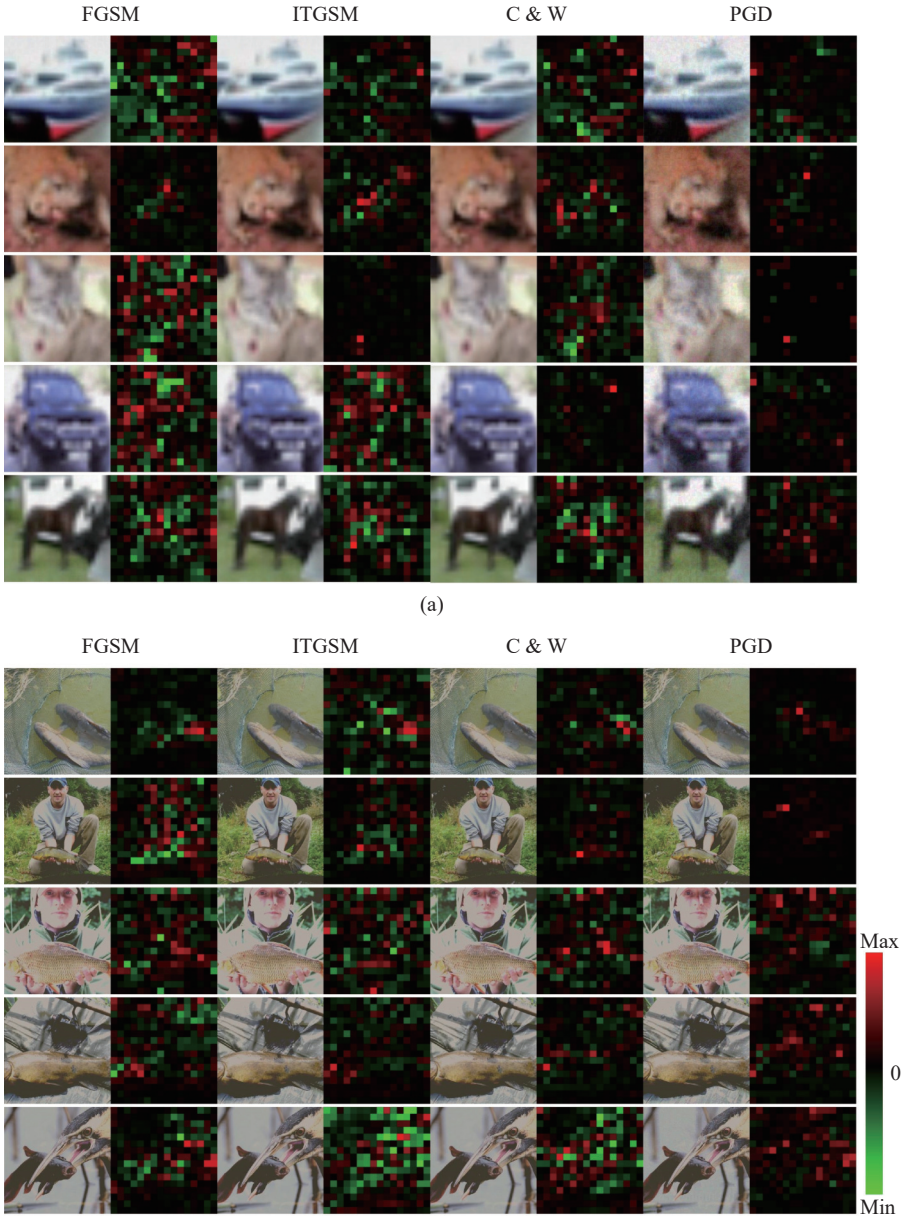

Fig. 9 Causal effects of different kinds of attacks: (a) Causal effects of the samples in CIFAR-10; (b) Causal effects of the samples in ImageNet. The clean images are shown in the first column. The adversarial samples generated by different algorithms and the corresponding causal effects are shown in the next columns.

Table 2 Quantitative results for different kinds of attaks

	CIFAR-10				ImageNet			
	mS	sdS	mTV	sdTV	mS	sdS	mTV	sdTV
FGSM target	0.872	0.075	72.40	28.09	0.812	0.079	76.25	36.86
FGSM untarget	0.887	0.095	68.55	33.24	0.814	0.091	71.93	35.66
ITFGSM target	0.894	0.056	75.12	32.61	0.818	0.077	80.45	38.31
ITFGSM untarget	0.886	0.092	68.51	33.87	0.804	0.097	65.13	34.49
C & W target	0.875	0.062	71.41	27.71	0.769	0.109	75.81	32.93
C & W untarget	0.916	0.058	66.97	36.26	0.850	0.064	61.54	35.88
PGD target	0.917	0.049	81.90	47.32	0.823	0.078	104.81	48.82
PGD untarget	0.935	0.061	72.05	33.39	0.826	0.089	79.84	28.33

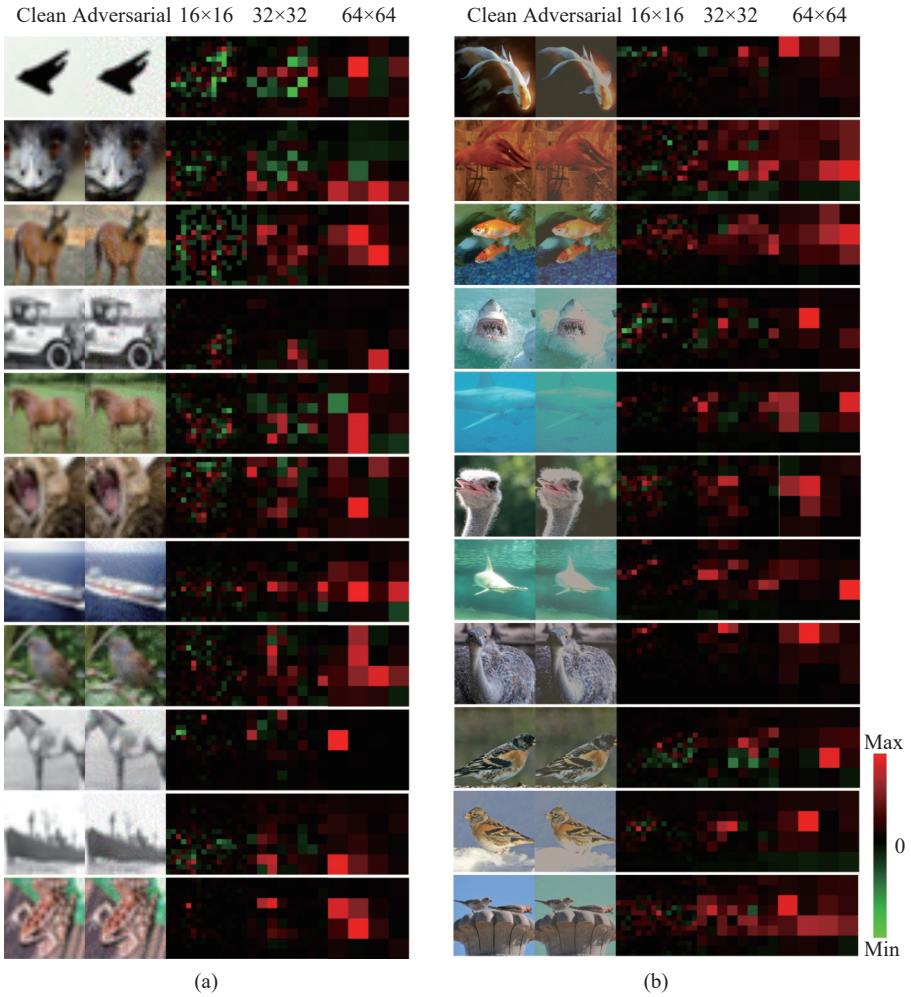

Fig. 10 Causal effects at different scales: (a) Causal effects of samples in CIFAR-10; (b) Causal effects of samples in ImageNet. The clean images are shown in the first column of (a) and (b), the adversarial samples are shown in the second column, and the causal effects at the three scales are shown in the other columns.

causal effects of most subregions have the opposite effect or are negligible. Different portions of the adversarial sample can be predicted as different labels by the recognition model. Hence, we can detect an adversarial attack according to the semantic inconsistency of the adversarial samples.

Specifically, for an input sample, we first obtain the predicted label by entering all of its subregions into the recognition model. For simplicity, this predicted label is called the global-label. Meanwhile, all the subregions of a sample are randomly split into k groups, where $2 < k < n$ and n is the number of subregions. Each group of subregions is input into the recognition model separately to obtain the predicted label of this group, which is named the partial-label. We regard this process as voting. If the winner, which is the partial-label with the most votes, is inconsistent with the global-label, the input sample is regarded as an adversarial sample; otherwise, the input sample is regarded as a clean sample. The pseudo-code for adversarial detection based on semantic inconsistency is shown in Algorithm 2.

Algorithm 2 Adversarial sample detection

Require: Input image: x; a division of x: $\{x_1, x_2, \cdots, x_n\}$; pre-trained ViT: $\mathcal{F}$; the number of groups: k.

Ensure: Prediction: $s \in \{0, 1\}$, $s = 0$ denotes that the input image is a clean sample, $s = 1$ denotes that the input image is an adversarial sample.

1) $s \leftarrow 0$

2) Obtain the global-label: $y_{global} \leftarrow \mathcal{F}(x)$

3) Divide $\{x_1, x_2, \cdots, x_n\}$ into k groups randomly: $\{\xi_1, \xi_2, \cdots, \xi_k\}$

4) **for** $i = 1 \rightarrow k$ **do**

5) Obtain the partial-label of ξ_i: $y^i_{partial} \leftarrow \mathcal{F}(\xi_i)$

6) **end for**

7) Obtain the winning partial label $y^*_{partial}$ from $\{y^1_{partial}, y^2_{partial}, \cdots, y^k_{partial}\}$

8) **if** $y_{global} \neq y^*_{partial}$ **then** $s \leftarrow 1$

9) **end if**

10) **return** s

Another instructive phenomenon is the inconsistency of causal effects at different scales, which indicates that

causal effects at different scales are complementary for adversarial attack detection. Hence, we propose a multi-scale strategy to combine the evidence at different scales. If the winning partial-label at any scale is inconsistent with the global-label, the input sample is regarded as an adversarial sample. The pseudo-code for adversarial detection based on multi-scale semantic inconsistency is shown in Algorithm 3.

Algorithm 3. Multi-scale adversarial sample detection

Require: Input image: x; pre-trained ViT: $\mathcal{F}$; the number of groups: k.

Ensure: Prediction: $s \in \{0, 1\}$, $s = 0$ denotes that the input image is a clean sample, and $s = 1$ denotes that the input image is an adversarial sample.

1) $s \leftarrow 0$

2) Split x at scale 16×16 (the size of each subregion is 16×16) as: $\{x_1, x_2, \cdots, x_{n16}\}$

3) Detect at scale 16×16:

$s_{16} \leftarrow$ Algorithm 2$(x, \{x_1, x_2, \cdots, x_{n16}\}, \mathcal{F}, k)$

4) **if** $s_{16} = 1$ **then** $s \leftarrow 1$

5) **end if**

6) Split x at scale 32×32 as: $\{x_1, x_2, \cdots, x_{n32}\}$

7) Detect at scale 32×32:

$s_{32} \leftarrow$ Algorithm 2$(x, \{x_1, x_2, \cdots, x_{n32}\}, \mathcal{F}, k)$

8) **if** $s_{32} = 1$ **then** $s \leftarrow 1$

9) **end if**

10) Split x at scale 64×64 as: $\{x_1, x_2, \cdots, x_{n64}\}$

11) Detect at scale 64×64:

$s_{64} \leftarrow$ Algorithm 2$(x, \{x_1, x_2, \cdots, x_{n64}\}, \mathcal{F}, k)$

12) **if** $s_{64} = 1$ **thten** $s \leftarrow 1$

13) **end if**

14) **return** s

5.2 Quantitative evaluation

To evaluate the proposed adversarial attack detection strategy, we conduct experiments on the CIFAR-10 and ImageNet test sets using MindSpore[77]. The test images consist of two kinds of samples: 1) all of the original images in the test images of CIFAR-10 and ImageNet, which are the negative samples for detection, and 2) the adversarial samples generated from the test images of CIFAR-10 and ImageNet by the attack method, which are the positive samples for detection. The negative and positive samples are balanced. Four kinds of adversarial attacks are explored: FGSM[13], C & W attack[20], PGD[21], and AutoAttack[78]. The adversary can perturb the clean sample within a certain amount:

$$\left\|x - x'\right\|_\infty < \epsilon \tag{6}$$

where $\epsilon = 8/255$ for all attack methods. The pre-trained ViT-Base/16 for image classification is adopted as the recognition model. All image samples are reshaped to 256×256. The number of groups is 4.

The performance is compared to three state-of-the-art adversarial attack detection approaches: DkNN[35], LID[33], and [34]. For comparison, the detection thresholds of these three methods are set to the threshold corresponding to the best accuracy. Their performances can be fully realized at this threshold since the negative and positive samples are balanced.

The results are shown in Tables 3 and 4. The results demonstrate the effectiveness of the proposed strategy for adversarial attack detection. According to the discoveries of causal inference, the proposed method outperforms the compared methods in most scenarios. For the adversarial training-based method LID, although it works pretty well in the WD scenario. There is an obvious gap between within dataset and cross dataset evaluation, which indicates that the generalization ability is quite limited. Note that the proposed method is not retrained by the adversarial samples or introduces an extra classifier, which means that we almost gain the capacity of adversarial attack detection for free. Meanwhile, multi-scale detection significantly increases the recall rates with a slight cost on the precisions, which demonstrates the complementar-

Table 3 Adversarial attack detection on CIFAR-10. The subscript LID_{WD} (within the dataset) means that LID is trained on the same attack it is evaluated on. The subscript LID_{CD} (cross dataset) means that the LID is trained on C & W adversarial examples, and tested on different unseen attacks.

Method	FGSM		FGSM		PGD		AutoAttack	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
DkNN[35]	63.74%	65.87%	60.11%	69.60%	57.33%	58.82%	58.37%	59.71%
LID_{WD}[33]	77.31%	79.82%	55.42%	54.12%	68.21%	65.70%	56.17%	57.86%
LID_{CD}[33]	73.98%	75.80%	–	–	61.01%	63.66%	55.53%	53.87%
Yu et al.[34]	87.21%	**88.20%**	91.51%	92.21%	61.48%	52.26%	55.27%	48.68%
Ours 16×16	98.04%	71.71%	**94.36%**	97.45%	86.98%	74.42%	**95.03%**	66.33%
Ours 32×32	98.17%	70.84%	94.06%	96.60%	87.06%	74.59%	90.31%	66.28%
Ours 64×64	**98.26%**	70.16%	94.03%	96.59%	**87.71%**	75.96%	87.09%	67.14%
Ours multi-scale	95.12%	85.02%	93.98%	**98.57%**	86.33%	**85.30%**	85.34%	**81.03%**

Table 4 Adversarial attack detection on ImageNet. Similar to Table 3, the subscript LID_{WD} and LID_{CD} refer to the within dataset and cross dataset evaluation, respectively.

Method	FGSM		C & W		PGD		AutoAttack	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
DkNN[35]	87.89%	86.27%	65.66%	61.01%	73.63%	79.79%	–	–
LID_{WD}[33]	**95.53%**	**96.85%**	53.26%	59.85%	**96.90%**	**95.23%**	–	–
LID_{CD}[33]	83.65%	88.30%	–	–	57.61%	58.14%	–	–
Yu et al.[34]	70.60%	74.53%	86.54%	84.41%	84.57%	89.42%	–	–
Ours 16×16	86.33%	84.95%	95.75%	97.26%	86.60%	85.80%	**82.44%**	79.95%
Ours 32×32	87.01%	86.08%	**96.24%**	97.84%	87.12%	86.61%	80.83%	80.99%
Ours 64×64	87.91%	87.63%	96.06%	97.35%	88.04%	88.02%	75.68%	82.99%
Ours multi-scale	85.61%	94.23%	95.21%	**98.61%**	85.63%	93.71%	75.01%	**90.96%**

ity of different scales.

5.3 Evaluation under adaptive attack

In order to critically evaluate the proposed method, we adopt an adaptive attack method to validate its performance in this subsection. The sparseness of the positive causal effect is the basis for the proposed detection method. A reasonable way for adaptive attacking is to increase the proportion of subregions with positive causal. Therefore, the different portions of an adversarial sample can lead to the same goal.

We conduct the adversarial attack on every small part of the input image, respectively, rather than the whole image. Moreover, the attacking objects are the same for all parts. Then all the parts are re-assembled together as an adversarial sample. Specifically, the input image is split into four parts by dividing all of its patches into four groups randomly. PGD is utilized to conduct the attack on all parts of the input image. The adversary can perturb the clean sample within 8/255.

The results are shown in Table 5. The performances of the proposed method under adaptive attack are passable. The results indicate that the proposed adversarial sample detection method is quite robust to adaptive attack. The underlying reason may be that it is difficult to unify the causal effects of different subregions of an image.

5.4 Sensitivity analysis on hyperparameter

There is a hyperparameter in the proposed method: the number of groups of patches k. To explore the impact of k on detection performance, we conduct a sensitivity analysis on it. The proposed method is tested when k variates from 2 to 6 under the attack of PGD.

The results are shown in Table 6. As the number of groups increases, the recall rates of detection increase. The percentage of groups that are affected by patches with positive causal decreases with the number of groups increasesing. Because these patches are sparse, there would be more partial-labels that are different with the goal of adversarial attacking. Hence, the recall rate increases with k. On the other hand, as the number of groups increases, the number of patches in each group will decrease. For clean samples, the prediction based on fewer patches would be more unstable, and the more clean samples would be wrongly determined as adversarial samples. Hence, the precision decreases with k. In summary, there is a trade-off between precision and recall rate in the selection of k.

Table 5 Evaluation under adaptive attacking

Scale	CIFAR − 10		ImageNet	
	Precision	Recall	Precision	Recall
16×16	**90.64%**	58.81%	**77.41%**	63.85%
32×32	86.70%	59.67%	75.87%	70.47%
64×64	78.50%	61.89%	75.51%	68.17%
Multi-scale	77.40%	**72.19%**	72.31%	**78.57%**

Table 6 Sensitivity analysis on the number of groups

k	CIFAR − 10		ImageNet	
	Precision	Recall	Precision	Recall
2	**88.87%**	84.28%	**88.75%**	92.57%
3	87.95%	84.73%	87.03%	93.06%
4	86.33%	85.30%	85.63%	93.71%
5	84.52%	85.95%	81.19%	94.46%
6	82.82%	**87.15%**	75.97%	**95.44%**

6 Recognition subregion based adversarial samples

Unlike from the general adversarial attack, which falsifies the whole image, subregion-based adversarial attack methods manipulate subregions of the image to fool the recognition model. There is usually no constraint on the

scale of adversarial perturbations. Hence, it is more difficult to recognize subregion-based adversarial samples. Meanwhile, a subregion-based adversarial attack is easy to realize physically, as shown in Fig. 11. These methods reduce the costs of adversarial attacks and increase the challenge for recognition systems in the real world.

According to the analysis based on causal inference, we propose a method for subregion-based adversarial sample recognition in this section. The proposed method is simple and effective, since causality provides us with a new viewpoint on the adversarial samples. Based on the proposed causal model, some crucial aspects of the working mechanism of adversarial samples are revealed, especially the sparseness of the positive causal effect. The causal perspective enables us to detect and recognize adversarial samples without extra models or training. Hence, the causal model and the causal effect estimation results are the preconditions of the proposed defense methods.

6.1 Recognition based on ensemble

The subregion-based adversarial attack is quite different from other kinds of adversarial attacks. The perturbations cannot be removed or handled like the other kinds of adversarial perturbations with minor scales. However, the analysis based on causal inference indicates that there is a common characteristic shared by the subregion-based and other adversarial samples: only a small portion of the image contributes to fooling the recognition model. Hence, we can recognize subregion-based adversarial samples based on this characteristic.

Specifically, a portion of the subregions is randomly selected and inputted into the recognition model to obtain the prediction. The above procedure is repeated k times to obtain k predicted labels. Since only a small portion of the image is manipulated, these subregions would not be sampled in most cases. Hence, we can recognize the subregion-based adversarial samples by ensembling the k predicted labels. The pseudo-code for the subregion-based adversarial sample recognition is shown in Algorithm 4.

Algorithm 4. Subregion-based adversarial sample recognition

Require: Input image: x; pre-trained ViT: $\mathcal{F}$; percentage of the sampled patches p; the number of sampling times: k.

Ensure: Predicted label: y.

1) **for** $i = 1 \rightarrow k$ **do**

2) Randomly sample patches from x, the percentage of the sampled patches is p:

$$\xi_i \leftarrow \{x_{i1}, x_{i2}, \cdots, x_{im}\}$$

3) Obtain the i-th predicted label: $y_i \leftarrow \mathcal{F}(\xi_i)$

4) **end for**

5) Obtain the winning final label y by voting:

$$y \leftarrow Vote(y_1, y_2, \cdots, y_k)$$

6) **return** y

Meanwhile, the causal analysis shows that the causal effects at different scales are inconsistent. Hence, similar to adversarial attack detection, we introduce the multi-scale strategy to combine the information of different scales.

6.2 Quantitative evaluation

To evaluate the proposed subregion-based adversarial sample recognition method, we conduct experiments on the test set of CIFAR-10 and ImageNet. The subregion-based adversarial attack method proposed by Komkov and Petiushko[78] is adopted to generate adversarial samples. The size of the manipulated subregion of CIFAR-10 is $\frac{1}{4} \times \frac{1}{4}$ of the input image, and the manipulated subregion of ImageNet is $\frac{1}{8} \times \frac{1}{4}$ of the input image. The pre-trained ViT-Base/16 for image classification is adopted as the recognition model. All image samples are reshaped to 256×256. The percentage of the sampled patches p is 25%, and the number of sampling times k is 5.

Two kinds of common defense strategies are considered for comparison: adversarial training and JPEG compression[47]. For the adversarial training strategy, the same recognition model is trained using the adversarial samples generated by FGSM. The quality of JPEG compression is 75.

The results are shown in Table 7. The results demonstrate the effectiveness of the proposed method. The recognition accuracy for the adversarial samples by our method is much better than the compared methods. The

Fig. 11 Example of subregion-based adversarial samples

Table 7 Recognition accuracy on clean and subregion-based adversarial samples

Method	CIFAR − 10			ImageNet		
	Clean ↑	Adversarial ↑	Gap ↓	Clean ↑	Adversarial ↑	Gap ↓
ViT	**95.67%**	9.80%	85.27%	**78.40%**	7.96%	70.44%
Adv. training	95.28%	11.19%	84.09%	72.76%	9.14%	63.62%
JPEG[33]	80.82%	75.68%	5.14%	59.72%	46.97%	12.75%
Ours 16 × 16	93.12%	89.90%	3.22%	69.55%	65.82%	3.73%
Ours 32 × 32	90.88%	86.30%	4.58%	70.29%	67.56%	2.73%
Ours 64 × 64	86.41%	79.01%	7.40%	68.31%	64.58%	3.73%
Ours multi-scale	93.20%	**90.28%**	**2.92%**	74.02%	**71.53%**	**2.49%**

gap between the performance of clean samples and adversarial samples by our method is minor. Note that we do not introduce any extra model or training data, which means that we almost significantly improve the robustness of the recognition model for free. Meanwhile, the performance of the multi-scale strategy is better than that of the single-scale strategy, which demonstrates the complementarity of different scales. The defense effect of adversarial training is weak, which indicates that the generalization ability of adversarial training is quite limited.

6.3 Sensitivity analysis on hyperparameter

There are two hyperparameters in our method: the percentage of sampled patches p and the number of sampling times k. To explore the impact of these hyperparameters on recognition performance, we conduct two experiments. The first experiment tests the recognition performance under different percentages of the sampled patches p. The number of sampling times k is 5 in this experiment. The second experiment tests the recognition performance under different numbers of sampling times k. The percentage of sampled patches p is 25% in this experiment.

The results are shown in Fig. 12. As we can observe, the recognition accuracy is highest when $p = 30\%$. When p is too small, there is not enough information for recognition. On the other hand, when p is too large, the adversarial region disturbs the recognition. For the number of sampling times, the recognition accuracy increases with k. However, the computational complexity also increases. Therefore, there is a trade-off between accuracy and efficiency.

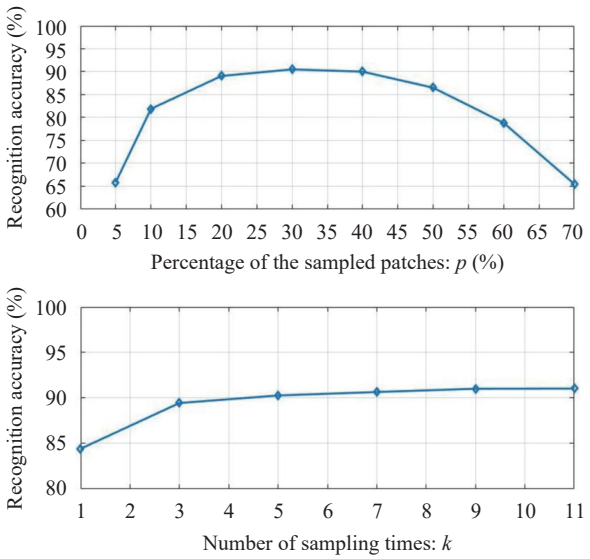

Fig. 12 A low sampling ratio (about 30%) works well for adversarial sample recognition. On the other hand, the recognition accuracy grows with k.

7 Conclusions

Although deep learning methods have made significant progress, they still lack effective and efficient defense strategies against adversarial attacks. As a result, the underlying working mechanism of adversarial samples has become the bottleneck of defense methods. In this paper, we adopt the methodology of causal inference to establish a causal model to describe the generation and performance of adversarial samples. Furthermore, this causal model enables us to attribute the output of the recognition model to the subregions of the input image and to interpret the working mechanism of adversarial samples. Hence, we can reveal many instructive phenomena of adversarial attacks and adversarial samples.

Based on the proposed causal model, we develop a method for adversarial attack detection and a method for adversarial sample recognition. These two methods are effective and efficient. Moreover, based on the powerful self-attention/transformers, we can detect and recognize adversarial samples without extra models or training. The results of the experiments demonstrate the superiority of our methods, especially the generalization capacity.

Acknowledgements

This work was supported by National Key Research and Development Program of China (No. 2020AAA

0140002), Natural Science Foundation of China (Nos. U1836217, 62076240, 62006225, 61906199, 62071468, 62176025 and U21B200389), and the CAAI-Huawei Mindspore Open Fund.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, USA, pp. 1106–1114, 2012.

[3] K. Simonyan, A Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA, 2015.

[4] C. Szegedy, W. Liu, Y. Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. Going deeper with convolutions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 1–9, 2015. DOI: 10.1109/CVPR.2015.7298594.

[5] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 2261–2269, 2017. DOI: 10.1109/CVPR.2017.243.

[6] J. Hu, L. Shen, G. Sun. Squeeze-and-excitation networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 7132–7141, 2018. DOI: 10.1109/CVPR.2018.00745.

[7] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[8] R. Girshick. Fast R-CNN. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Santiago, Chile, pp. 1440–1448, 2015. DOI: 10.1109/ICCV.2015.169.

[9] J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 779–788, 2016. DOI: 10.1109/CVPR.2016.91.

[10] J. Long, E. Shelhamer, T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 3431–3440, 2015. DOI: 10.1109/CVPR.2015.7298965.

[11] K. M. He, G. Gkioxari, P. Dollár, R. Girshick. Mask R-CNN. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 2980-2988, 2017. DOI: 10.1109/ICCV.2017.322.

[12] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations*, Banff, Canada, 2014.

[13] I. J. Goodfellow, J. Shlens, C. Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA, 2015.

[14] S. M. Moosavi-Dezfooli, A. Fawzi, P. Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 2574–2582, 2016. DOI: 10.1109/CVPR.2016.282.

[15] F. Schroff, D. Kalenichenko, J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 815–823, 2015. DOI: 10.1109/CVPR.2015.7298682.

[16] M. Ren, Y. L. Wang, Z. N. Sun, T. N. Tan. Dynamic graph representation for occlusion handling in biometrics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Palo Alto, USA, pp. 11940–11947, 2020. DOI: 10.1609/aaai.v34i07.6869.

[17] M. Ren, C. Y. Wang, Y. L. Wang, Z. N. Sun, T. N. Tan. Alignment free and distortion robust iris recognition. In *Proceedings of International Conference on Biometrics*, IEEE, Crete, Greece, 2019. DOI: 10.1109/ICB45273.2019.8987369.

[18] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. N. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. F. Chen, D. Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 2443–2451, 2020. DOI: 10.1109/CVPR42600.2020.00252.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[20] N. Carlini, D. Wagner. Towards evaluating the robustness of neural networks. In *Proceedings of IEEE Symposium on Security and Privacy*, IEEE, San Jose, USA, pp. 39–57, 2017. DOI: 10.1109/SP.2017.49.

[21] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

[22] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard. Universal adversarial perturbations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 86–94, 2017. DOI: 10.1109/CVPR.2017.17.

[23] C. N. Zhang, P. Benz, A. Karjauv, I. S. Kweon. Data-free universal adversarial perturbation and black-box attack. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 7848–7857, 2021. DOI: 10.1109/ICCV48922.2021.00777.

[24] Z. B. Wang, H. C. Guo, Z. F. Zhang, W. X. Liu, Z. Qin, K.

Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 7619–7628, 2021. DOI: 10.1109/ICCV48922.2021.00754.

[25] Z. Yuan, J. Zhang, Y. P. Jia, C. Q. Tan, T. Xue, S. G. Shan. Meta gradient adversarial attack. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 7728–7737, 2021. DOI: 10.1109/ICCV48922.2021.00765.

[26] J. W. Su, D. V. Vargas, K. Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019. DOI: 10.1109/TEVC.2019.2890858.

[27] A. Athalye, L. Engstrom, A. Ilyas, K. Kwok. Synthesizing robust adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, pp. 284–293, 2018.

[28] T. B. Brown, D. Mané, A. Roy, M. Abadi, J. Gilmer. Adversarial patch. [Online], Available: https://arxiv.org/abs/1712.09665, 2017.

[29] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. W. Xiao, A. Prakash, T. Kohno, D. Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 1625–1634, 2018. DOI: 10.1109/CVPR.2018.00175.

[30] M. Sharif, S. Bhagavatula, L. Bauer, M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, ACM, Vienna, Austria, pp. 1528–1540, 2016. DOI: 10.1145/2976749.2978392.

[31] K. D. Xu, G. Y. Zhang, S. J. Liu, Q. F. Fan, M. S. Sun, H. G. Chen, P. Y. Chen, Y. Z. Wang, X. Lin. Adversarial t-shirt! evading person detectors in a physical world. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 665–681, 2020. DOI: 10.1007/978-3-030-58558-7_39.

[32] R. Feinman, R. R. Curtin, S. Shintre, A. B. Gardner. Detecting adversarial samples from artifacts. [Online], Available: https://arxiv.org/abs/1703.00410, 2017.

[33] X. J. Ma, B. Li, Y. S. Wang, S. M. Erfani, S. N. R. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, J. Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

[34] T. Yu, S. Y. Hu, C. Guo, W. L. Chao, K. Q. Weinberger. A new defense against adversarial images: Turning a weakness into a strength. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 1633–1644, 2019.

[35] N. Papernot, P. McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. [Online], Available: https://arxiv.org/abs/1803.04765, 2018.

[36] K. Lee, K. Lee, H. Lee, J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 7167–7177, 2018.

[37] A. Kurakin, I. J. Goodfellow, S. Bengio. Adversarial machine learning at scale. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017.

[38] T. Na, J. H. Ko, S. Mukhopadhyay. Cascade adversarial machine learning regularized with a unified embedding. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

[39] F. Tramèr, A. Kurakin, N. Papernot, I. J. Goodfellow, D. Boneh, P. D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

[40] A. S. Ross, F. Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence Conference and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, USA, pp. 203, 2018.

[41] G. Cazenavette, C. Murdock, S. Lucey. Architectural adversarial robustness: The case for deep pursuit. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 7146–7154, 2021. DOI: 10.1109/CVPR46437.2021.00707.

[42] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Proceedings of IEEE Symposium on Security and Privacy*, IEEE, San Jose, USA, pp. 582–597, 2016. DOI: 10.1109/SP.2016.41.

[43] X. Y. Cao, N. Q. Gong. Mitigating evasion attacks to deep neural networks via region-based classification. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, ACM, Orlando, USA, pp. 278–287, 2017. DOI: 10.1145/3134600.3134606.

[44] H. Lee, S. Han, J. Lee. Generative adversarial trainer: Defense to adversarial perturbations with GAN. [Online], Available: https://arxiv.org/abs/1705.03387, 2017.

[45] Y. Jang, T. C. Zhao, S. Hong, H. Lee. Adversarial defense via learning to generate diverse attacks. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Korea, pp. 2740–2749, 2019. DOI: 10.1109/ICCV.2019.00283.

[46] M. Moayeri, S. Feizi. Sample efficient detection and classification of adversarial attacks via self-supervised embeddings. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 7657–7666, 2021. DOI: 10.1109/ICCV48922.2021.00758.

[47] N. Das, M. Shanbhogue, S. T. Chen, F. Hohman, L. Chen, M. E. Kounavis, D. H. Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression. [Online], Available: https://arxiv.org/abs/1705.02900, 2017.

[48] T. Salimans, A. Karpathy, X. Chen, D. P. Kingma. PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. In *Proceed-*

ings of the 5th International Conference on Learning Representations, Toulon, France, 2017.

[49] Y. Song, T. Kim, S. Nowozin, S. Ermon, N. Kushman. PixelDefend: Leveraging generative models to understand and defend against adversarial examples. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

[50] Y. Bai, Y. Feng, Y. S. Wang, T. Dai, S. T. Xia, Y. Jiang. Hilbert-based generative defense for adversarial examples. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Korea, pp. 4783–4792, 2019. DOI: 10.1109/ICCV.2019.00488.

[51] S. M. Moosavi-Dezfooli, A. Shrivastava, O. Tuzel. Divide, denoise, and defend against adversarial attacks. [Online], Available: https://arxiv.org/abs/1802.06806, 2018.

[52] B. Sun, N. H. Tsai, F. C. Liu, R. Yu, H. Su. Adversarial defense by stratified convolutional sparse coding. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 11439–11448, 2019. DOI: 10.1109/CVPR.2019.01171.

[53] D. W. Zhou, N. N. Wang, C. L. Peng, X. B. Gao, X. Y. Wang, J. Yu, T. L. Liu. Removing adversarial noise in class activation feature space. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 7858–7867, 2021. DOI: 10.1109/ICCV48922.2021.00778.

[54] J. Pearl, M. Glymour, N. P. Jewell. *Causal Inference in Statistics: A Primer*, Chichester, UK: John Wiley & Sons, 2016.

[55] B. Scholkopf. Causality for machine learning. *Probabilistic and Causal Inference*: *The Works of Judea Pearl*, pp. 765–804, 2022.

[56] K. Chalupka, P. Perona, F. Eberhardt. Visual causal feature learning. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, Amsterdam, Netherlands, pp. 181–190, 2015.

[57] T. Wang, J. Q. Huang, H. W. Zhang, Q. R. Sun. Visual commonsense R-CNN. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 10757–10767, 2020. DOI: 10.1109/CVPR42600.2020.01077.

[58] Z. Q. Yue, H. W. Zhang, Q. R. Sun, X. S. Hua. Interventional few-shot learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 2734–2746, 2020.

[59] K. H. Tang, J. Q. Huang, H. W. Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 1513–1524, 2020.

[60] D. Zhang, H. W. Zhang, J. H. Tang, X. S. Hua, Q. R. Sun. Causal intervention for weakly-supervised semantic segmentation. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 655–666, 2020.

[61] L. Chen, X. Yan, J. Xiao, H. W. Zhang, S. L. Pu, Y. T. Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 10797–10806, 2020. DOI: 10.1109/CVPR42600.2020.01081.

[62] W. J. Zhang, L. Liu, J. Y. Li. Robust multi-instance learning with stable instances. In *Proceedings of the 24th European Conference on Artificial Intelligence*, Santiago de Compostela, Spain, pp. 1682–1689, 2020.

[63] C. Wang, X. T. Lu, W. Wang. A theoretical analysis based on causal inference and single-instance learning. *Applied Intelligence*, to be published. DOI: 10.1007/s10489-022-03193-0.

[64] H. Hu, Z. Zhang, Z. D. Xie, S. Lin. Local relation networks for image recognition. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Korea, pp. 3463–3472, 2019. DOI: 10.1109/ICCV.2019.00356.

[65] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens. Stand-alone self-attention in vision models. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 68–80, 2019.

[66] H. S. Zhao, J. Y. Jia, V. Koltun. Exploring self-attention for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 10073–10082, 2021. DOI: 10.1109/CVPR42600.2020.01009.

[67] A. Srinivas, T. Y. Lin, N. Parmar, J. Shlens, P. Abbeel, A. Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 16514–16524, 2021. DOI: 10.1109/CVPR46437.2021.01625.

[68] J. Y. Gu, H. Hu, L. W. Wang, Y. C. Wei, J. F. Dai. Learning region features for object detection. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 392–406, 2018. DOI: 10.1007/978-3-030-01258-8_24.

[69] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 10347–10357, 2021.

[70] L. Yuan, Y. P. Chen, T. Wang, W. H. Yu, Y. J. Shi, Z. H. Jiang, F. E. H. Tay, J. S. Feng, S. C. Yan. Tokens-to-token ViT: Training vision transformers from scratch on ImageNet. In *Proceedings of IEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 538–547, 2021. DOI: 10.1109/ICCV48922.2021.00060.

[71] X. X. Chu, B. Zhang, Z. Tian, X. L. Wei, H. X. Xia. Do we really need explicit position encodings for vision transformers? [Online], Available: https://arxiv.org/abs/2102.10882, 2021.

[72] J. Pearl. Direct and indirect effects. *Probabilistic and Causal Inference*: *The Works of Judea Pearl*, pp. 373–392, 2022.

[73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA, pp. 6000–6010, 2017.

[74] A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Citeseer, 2009.

[75] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Miami, USA, pp. 248–255, 2009. DOI: 10.1109/CVPR.2009.5206848.

[76] A. Kurakin, I. J. Goodfellow, S. Bengio. Adversarial examples in the physical world. *Artificial Intelligence Safety and Security*, R. V. Yampolskiy, Ed., New York, USA: Chapman and Hall/CRC, pp. 1–14, 2018.

[77] F. Croce, M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 2206-2216, 2020.

[78] S. Komkov, A. Petiushko. AdvHat: Real-world adversarial attack on ArcFace face ID system. In *Proceedings of the 25th International Conference on Pattern Recognition*, IEEE, Milan, Italy, pp. 819–826, 2021. DOI: 10.1109/ICPR48806.2021.9412236.

Min Ren received the B. Eng. degree in mechanical engineering and automation from National University of Defense Technology, China in 2013. Currently, he is a Ph. D. degree candidate with School of Artificial Intelligence, University of Chinese Academy of Sciences, China, and Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China.

His research interests include pattern recognition, computer vision and biometrics.

E-mail: min.ren@cripac.ia.ac.cn

ORCID: 0000-0002-0126-1726

Yun-Long Wang received the M. Sc. and Ph. D. degrees in pattern recognition and intelligent systems from Department of Automation, University of Science and Technology of China, China. He is currently an associate professor with CRIPAC, NLPR, CASIA, China.

His research interests include pattern recognition, machine learning, light-field photography, and biometrics.

E-mail: yunlong.wang@cripac.ia.ac.cn (Corresponding author)

ORCID: 0000-0002-3535-308X

Zhao-Feng He received the Ph. D. degree in pattern recognition and intelligent systems from CASIA, China in 2010. Currently, he is a professor at Beijing University of Posts and Telecommunications (BUPT) and is the founder of the Laboratory of Visual Computing and Intelligent System (VCIS), China.

His research interests include biometrics, computer vision, and intelligent system.

E-mail: zhaofenghe@bupt.edu.cn